

Matching Phrases for Arabic-to-English Example-Based Translation

Kfir Bar

School of Computer Science
Tel Aviv University
Ramat Aviv, Israel

kfirbar@post.tau.ac.il

Yaacov Choueka

School of Computer Science
Bar-Ilan University
Ramat Gan, Israel

ycesarah@netvision.net.il

Nachum Dershowitz

School of Computer Science
Tel Aviv University
Ramat Aviv, Israel

nachumd@cs.tau.ac.il

Abstract

An implementation of a non-structural Example-Based Machine Translation system that translates sentences from Arabic to English, using a parallel corpus aligned at the paragraph level, is described. Each new input sentence is fragmented into phrases and those phrases are matched to example patterns, using various levels of morphological data. The system has been implemented and automatically evaluated. Results are encouraging.

1 Introduction

Ever since it was first proposed by Nagao (1984), the example-based (or “memory-based”) paradigm has become a fairly common technique for natural language processing (NLP), and especially for machine-translation applications. The main idea behind example-based machine translation (EBMT) is to translate fragments of the source-language input text based on similar known example translations. Such a process presumably emulates the way a human translator operates in some cases. Since it uses real human-translated data, the resultant translations are usually more fluent than ones created artificially using other translation paradigms.

We have developed an Arabic-to-English example-based translation system for short sentences. It exploits a bilingual corpus to find examples that match fragments of the input source-language (Modern Standard Arabic-MSA, in our case) text,

and imitates its translations. Here we show how the results are improved by forcing the matching step to work on the phrase level only. The operant definition of a phrase for us is a combination of adjacent base-phrases of the input sentence. Base-phrases are extracted using the AMIRA (Diab et al., 2004) tool, as will be described.

In the transfer step, those matched phrases are translated using the target-language (English, in our case) version of the parallel corpus. In the recombination step of an example-based translation system, all the translated fragments are pasted together to form a complete target-language text.

Like many other Semitic languages, Arabic is highly inflected; words are derived from a *root* and *pattern* (the stem), combined with prefixes, suffixes and circumfixes. The root consists of 3-4 consonants and the pattern is a sequence of consonants and variables. Using the same root with different patterns may yield a word with a different meaning. For instance, the combination of the root ك.ت.ب (*k.t.b*) and the pattern mXXX (here, X is a variable) results in the word مكتب (*mktb*, “office”). Combining the same root with the pattern XXAX, results in the word كتاب (*ktAb*, “book”).

In working with a highly inflected language, finding an exact match for an input phrase with reasonable precision presumably requires a large parallel corpus. Unlike other implementations, our system uses only a small parallel corpus for translation. Matching phrases to the corpus is done on different linguistic levels, so that not only exact phrases are discovered but also related ones.

The system described here is a non-structural system, so it stores the translation examples as textual strings, with some additional linguistic features. Our work on this system is still in progress; currently, it translates each phrase separately and concatenates those translations to form an output target-language sentence. Recombining those translations into a final, coherent form is left for future work.

The following section gives short description of some previous works. Section 3 is a general description of our system. In Section 4, we present some experimental results using common automatic metrics. Conclusions are presented in Section 5.

2 Previous Work

The initiator of the example-based approach applied to machine-translation is Nagao (1984), who investigated a structural Japanese-to-English example-based system. Other influential works include (Sato and Nagao, 1990; Maruyama and Watanabe, 1992; Sumita and Iida, 1995; Nirenberg et al. 1994; Brown, 1999). Recent, independent work on Arabic-to-English is (Philips et al., 2007). That system uses morphological-generalization to better exploit the bilingual corpus for matching.

Our work focuses on reducing the number of matched fragments for the transfer step, as well as on handling various issues regarding the specific pair Arabic-English. In contrast with other implementations, our system uses only a relatively small corpus.

3 System Description

3.1 Translations Corpus

The translation examples in our system were extracted from a collection of parallel unvocalized Arabic-English documents taken from the United-Nations document inventory, available under the Official-Document-System (ODS). The documents were automatically aligned on the paragraph level, and each parallel paragraph was taken as a translation example. Based on our final goal of creating a translation-corpus, we could afford using a imperfect alignment algorithm, since we could later use only those paragraphs that were successfully aligned with high confidence. We used a version of

the algorithm for Hebrew-English pair described in Choueka et al. (2000), which is based on the DK-Vec algorithm (Fung and McKeown, 1994). Fortunately, most of the UN documents are divided into paragraphs delimited by the new-line character, so we basically used the alignment algorithm to find as many *word-anchors* as possible and used them to find matches for each Arabic paragraph. By “word-anchor” we mean an Arabic word whose equivalent is discovered in the English version of the paragraph.

All the translation examples were morphologically analyzed using the well-known Buckwalter morphological analyzer (version 1.0) (Buckwalter, 2002), and then part-of-speech tagged using AMIRA in such a way that, for each word, we considered only the relevant Buckwalter analyses with the corresponding AMIRA part-of-speech tag. For each Arabic word in the translation example, we look up its English equivalents in a lexicon, created using the Buckwalter glossaries, and then expand those English words with synonyms from WordNet. Then we search the English version of the translation example for all instances on the lemma level and insert them in a special lookup table.

The Arabic version of the corpus was indexed on the word, stem and lemma levels (stem and lemma, as defined by the Buckwalter analyzer), so, for each given word, we are able to retrieve all translation examples that contain that word on any of those three levels.

3.2 Matching Phrases

Given a new input sentence, the system begins by searching the corpus for translation examples for which the Arabic version matches fragments of the input sentence. In the implementation we are describing, the system is restricted to fragmenting the input sentence so that a matched fragment must be a combination of one or more complete adjacent base-phrases of the input sentence. The base-phrases are initially extracted using the AMIRA tool. Fragments also must contain at least two words. For instance, take the following sentence:

يكون المعهد قادرا على القيام ببحوث مستقلة

(Ykwn AlmEhd qAdrA EIY AlqyAm bbHwv mstqlp, "The institute is able to pursue independent research"). Its AMIRA base-phrases are:

[VP ykwn] [NP AlmEhd] [ADJP qAdrA]
 [PP EIY AlqyAm] [PP bbHwv mstqlp]

That means, for example, that the fragment ykwn AlmEhd qAdrA is possible, but the fragment EIY AlqyAm bbHwv is not allowed, because it is not a combination of complete adjacent base-phrases. Note that matching the complete input sentence is allowed. Currently, we have not taken the types of base-phrases into consideration, but it seems that using this kind of information, compiled into several pattern rules (e.g. matching the sequence PP NP), will improve the matching results, by forcing the system to only consider reasonable sequences of base-phrases.

The same fragment can be found in more than one translation example. Therefore, a “match-score” is assigned to each fragment-translation pair, signifying the quality of the matched fragment in the specific translation example.

Fragments are matched word by word, so the score for a fragment is the average of the individual word match-scores. Words are matched on text, stem, lemma, and morphological levels, with each level assigned a different score. Text (exact string) and stem matches credit the words with the maximum possible. The lemma of a word is revealed using the Buckwalter analyzer, and matching words on that level credits them with fewer points. The morphological level credits the fragment match-score with a minimal amount. Table 1 summarizes the ad-hoc match level scores we used in these experiments.

Text and stem matches receive almost the same score, since, currently, we do not yet handle the translation modification needed. When dealing with unvocalized text, there are, of course, complicated situations when both words have the same unvocalized stem but different lemmas, for example, the words كتب (katab, “wrote”) and كتب (kutub, “books”). Such cases are not yet handled accurately, since we have not worked with a context-sensitive Arabic lemmatizer and so cannot derive the correct lemma of an Arabic word. Still, the combination of the Buckwalter morphological analyzer and the AMIRA part-of-speech tagger allows us to reduce the number of possible lemmas for every Arabic word so as to reduce the amount of ambiguity. Actually, by “lemma match”, we mean that words match on any one of their possible lemmas. The match-score in such a case is determined by the ratio between the number of equal

lemmas and the total number of lemma pairs (one per word). Further investigation, as well as working with a context-sensitive morphology analyzer (Habash and Rambow, 2005), will allow us to better handle all such situations.

Match Level	Description	Match Score
<i>Text</i>	Exact match of words.	1.0
<i>Stem</i>	Stems of words match. For instance, the words الدستورية (<i>Aldstwryp</i> , “the constitutionality”) and الدستوري (<i>dstwryty</i> , “my constitutional”) share the stem دستوري (<i>dusotuwryy</i>)	0.9
<i>Lemma</i>	Words share a lemma. For instance, the following words match in their lemmas, but not stems: مارق (<i>mAriq</i> , “apostate”); مراق (<i>mur-Aq</i> , “apostates”)	dynamic
<i>Semantic</i>	This level is planned but not yet implemented. The idea is that, for example, two location names would get a higher score than two dissimilar proper nouns.	0.8
<i>Morphology</i>	Words match only in their part-of-speech (e.g. both are nouns). We also require that both have the same tags for affixes. For example, if a noun has the definite-article prefix ال (<i>Al</i> , “the”), the matched word must be a noun and also have the definite article prefix.	0.3
<i>Common Word</i>	This level includes common words and affixes from a predefined list. These are organized in groups with the same meaning. (Clearly, a word/affix may be a member of more than one group.) Members of the same group are also matched on this level. For example the prefix ب (<i>b</i> , “with”, “by”, “in”) is in the same group as the preposition في (<i>fj</i> , “in”).	1.0

Table 1. Word-matching levels

Restricting the system to work only with combinations of base-phrases is used for reducing the number of fragments for the transfer stage, where they are each translated. Table 2 gives the amount of reduction. The left column’s value is the average number of fragments that were matched per input sentence, without restricting the system to match on phrases. This evaluation was performed on 300 input sentences. The right column’s value is the same average for the same 300 input sentences using the phrase restriction.

Average number of fragments without Base-Phrase Restriction	Average number of fragments number with Base-Phrase Restriction
1824.08	655.02

Table 2. Comparing fragments

Fragments with a score below some predefined threshold are discarded, since passing low-score

fragments to the next step dramatically increases total running time and it was sometimes unfeasible to process all those fragments. Note that a larger corpus, with the concomitant increase in the number of potential fragments, would require raising the threshold.

Fragments are stored in a structure comprising the following: (1) source pattern – fragment’s Arabic text, taken from the input sentence; (2) example pattern – fragment’s Arabic text, taken from the matched translation example; (3) example – the English translation of the example pattern; (4) match score – of the fragment and its example translation. For efficiency, fragments sharing the same example pattern are collected and stored in a higher-level, “general-fragment” structure. (A general-fragment consisting of only one fragment is possible.) All general-fragments composed of fragments that successfully passed the threshold are moved to the next step – the transfer step.

3.3 Transfer

The input to the transfer step consists of all the collected general-fragments found in the matching step, and the output is the translations of those general-fragments. The translation of a general-fragment is taken to be the best generated translation among the comprised fragments. Translating a fragment is done in two main steps: (1) extracting the translation of the example pattern from the English version of the translation example; (2) fixing the extracted translation so that it will be the translation of the fragment’s source pattern.

First Step – Translation Extraction

The first step is to extract the translation of the fragment’s example pattern from the English version of the translation example. Here we use the prepared look-up table for every translation example within our corpus. For every Arabic word in the pattern, we look up its English equivalents in the table and mark them in the English version of the translation example. Then, we extract the shortest English segment that contains the maximum number of equivalence words. Usually a word in some Arabic example pattern has several English equivalents, which makes the translation extraction process complicated and error prone. For this reason, we also restrict the ratio between the number of Arabic words in the example pattern

and the number of English words in the extracted translation, bounding them by a function of the ratio between the total number of words in the Arabic and English versions of the translation example.

For example, take the following translation example:

A: الخدمات الاستشارية والتعاون التقني في ميدان حقوق الإنسان

E: “Advisory services and technical cooperation in the field of human rights.”

Table 3 is the corresponding look-up table. Now, suppose the example pattern is ميدان حقوق الإنسان (mydAn Hqwq Al<nsAn, “the field of human rights”), so we want to extract its translation from the English version of the translation example. Using the extracted look-up, we mark the English equivalences of the pattern words in the translation example: “Advisory services and technical cooperation in the field of human rights”, and then we extract the shortest English segment that contains the maximum number of equivalent words, viz. “field of human rights”.

English	Arabic
Services	الخدمات
Advisory	الاستشارية
Cooperation	والتعاون
Technical	التقني
In	في
Field	ميدان
Rights	حقوق
Human	الإنسان

Table 3. Alignment look-up table

This is, of course, a simple instance. More complicated ones would have more than one equivalent per Arabic word.

Sometimes it is hard to find the corresponding English equivalents for a specific Arabic word. Usually this happens when the Arabic word is part of some phrase, whereas its translation does not follow word for word, as in, for example, the Arabic example pattern غير رسمي (gyr rsm), meaning “not formal”. In many cases, we might find “informal” in the English version instead. The problem is that neither the synonym list of the word رسمي (rsm, “formal”), nor the list of the word غير (gyr, “not”), contains the word “informal”. Such a situation is handled by a manually defined rule that is triggered whenever the word غير (gyr, “not”) appears. The system checks the

following word, and -- instead of building a synonym list -- builds an antonym list, using WordNet. In this example, the word “informal” appears as an antonym of the “formal” in WordNet.

There are more complicated structures that are not handled yet, but capturing and writing rules for such cases seems quite feasible.

Second Step – Fixing the Translation

Recall that the match of a corpus fragment to the input fragment can be inexact: words may be matched on several levels. Exactly matched words are assumed to have the same translation, but stem or lemma matched words may require modifications (mostly inflection and prepositions issues) to the extracted translation. These issues are still left for future work. Words matched on the part-of-speech level require complete change of meaning. For example, take the input fragment مجلس الامن (mjls AlAmn, “the Security Council”), matched to the fragment مسؤولية الامن (ms&wlyp AlAmn, “the security responsibility”) in some translation example. The words مجلس (mjls, “council”) and مسؤولية (ms&wlya, “responsibility”) are matched on the part-of-speech level (both are nouns). Assume that the extracted translation from the translation example is “the security responsibility”, which is actually a translation of مسؤولية الامن (ms&wlyp AlAmn, “the security responsibility”) and is not the translation of the input pattern at all. But, by replacing the word “responsibility” from the translation example with the translation of مجلس (mjls, “council”) from the lexicon, we get the correct phrase: “the security council”. The lexicon is implemented using the glossaries extracted from the Buckwalter morphological analyzer and expanded with WordNet synonyms as was explained above.

Sometimes the extracted translation contains some extra unnecessary words in the middle. Those words appear mostly because of the different structure of a noun-phrase in both languages. For example, consider the example, موضوع الامن الاقليمي (mwDwE AlAmn AlAqlymy), and its translation: “the subject of regional security”. By extracting the translation of the pattern موضوع الامن (mwDwE AlAmn), we obtain: “the subject of regional security” (since it is the shortest segment that contains maximum word alignments). Clearly, the word “regional” is unnecessary in the translation because it is the translation of the word الاقليمي (AlAqlymy, “the regional”) that does not appear in

the pattern. So by removing that word from the translation we obtain the correct translation of the pattern. The word “regional” appears in the extracted translation due to the fact that Arabic adjectives come after the nouns they qualify, which is the opposite of English syntax. Here, the noun-phrase الامن الاقليمي (AlAmn AlAqlymy, “the regional security”) is translated so that the translation of الاقليمي (AlAqlymy, “the regional”) appears before the translation of الامن (AlAmn, “security”). Currently, identifying such situations is done by searching for the translation of the word “regional” in a fixed number of Arabic words that come immediately after the pattern in the translation example. However, this method is insufficient for more complex situations and is also very time consuming. Our plan is to use the boundaries of the noun-phrase, given by AMIRA, to delimit the search area.

Removing unnecessary words from the extracted translation must preserve the correct English syntax of the remaining translation, which in some cases seems to be a difficult task. For that purpose, we have compiled several rules to deal with different situations. These rules are based on the syntax of the English extracted translation and identify cases that need special care. First, we chunk the translation to discover its basic noun-phrases, using the BaseNP (Ramshaw and Marcus, 1995) chunker. To do that, we first apply Brill’s part-of-speech tagger (Brill, 1992) to the translation. Then, by looking at the chunked English text, we can ascertain the effect of removing the unnecessary word. In the previous example, removing the word “regional” from the text, “the subject of regional security”, may be done without any further modification, since by tagging and chunking the segment we get:

[the/DT subject/NN] of/IN
[regional/JJ security/NN]

(the phrases in brackets are noun-phrases) and “regional” is simply an adjective within a noun-phrase, which still has the same head. Prepositions and other function-words that relate to the phrase are still necessary, so we keep them.

As already mentioned, a general-fragment may contain several fragments sharing the same Arabic example pattern. Among the extracted translations of the comprised fragments, which are all translations of the same Arabic pattern, we choose the translation that covers the maximum number of

Arabic words to represent the general-fragment. The translation-score calculated for the chosen translation is the ratio between the number of covered words and the total number of words in the Arabic pattern. The total-score of a general-fragment is the multiplication of its match-score and its translation-score.

3.4 Recombination

In the recombination step, we paste together the extracted translations to form a complete translation of the input sentence. This is generally composed of two subtasks. The first is finding the N best recombinations of the extracted translations that cover the entire input sentence, and the second is smoothing out the recombined translations to make a fully grammatical English sentence. Currently, we handle only the first subtask. Although the second task is left for future work, we observed that by matching on base-phrase combinations, we have actually improved the recombined translations, over the previous implementation, even if it was not smoothed out. That is reasonable since the fragments – the translated base-phrase combinations – combine in a better way than translations of fragments that were extracted without any restrictions.

By multiplying the total-scores of the comprised general-fragments, we calculate a final-translation-score for each generated recombination. The N best (where N is a configurable parameter) recombinations are reported. Another observation here is that the number of the N best results is usually less than that number when working without the base-phrase restriction, as can be seen in Table 4 when translating 300 sentences separately.

Average number of N best final translations without Base-Phrase Restriction	Average number of N best final translations with Base-Phrase Restriction
48.3044	2.4298

Table 4. Comparing N best

4 Experimental Results

Experiments were conducted on a corpus containing 16,500 translation examples only. The following results are based on a test set of 300 Arabic short sentences (6 words per sentence, on average),

taken from unseen documents of the United-Nations inventory.

Experiment	BLEU	NIST	METEOR
No BP (B)	0.2167	4.0574	0.4761
BP (B)	0.2209	4.2636	0.4147
No BP (R)	0.2588	5.0231	0.5295
BP (R)	0.3026	5.2324	0.5055

Table 5. Experimental Results: No BP (not using the Base-Phrase restriction); BP (using Base-Phrase restriction); B (system's best result); R (best result chosen by a human referee)

Despite the fact that our system still does not perform the last, smoothing stage of the translation process and is still under construction, we evaluated its ten best results under some of the common automatic criteria for machine-translation evaluation: BLEU (Papineni, 2002), NIST, and METEOR (Banerjee and Lavie, 2005). Also, we used only two different translation references for the evaluation. Table 5 shows some experimental results. The first two rows contain the results of evaluating the system's highest ranked translation for each input sentence, with and without the restriction for matching only sequences of base-phrases. The last two rows are the same, but on the best translation from the viewpoint of a human referee. In most cases, the best translation chosen by the referee had a close (or even the same) final-translation-score as the system's best translation. From the results we can observe that in spite of the small reduction in METEOR, the system performs better when matching sequences of complete base-phrases of the input sentence. Another important observation is the increase in the difference of BLEU between the last two rows as opposed to the same difference between the first two lines. That means, that even though the amount of N best translations was reduced when the system uses the base-phrase restriction, the highest ranked translation is not always the best.

5 Conclusions

The system we are working on has demonstrated the potential for example-based approach for Arabic, with only minimum investment in Arabic syntactical and linguistic issues. We found that restricting the system to keep only those fragments that form a sequence of complete base-phases re-

sults in better translations and reduces computational effort. Considering the type of those phrases is definitely a direction for future investigation. Matching fragments on the level of lemma and stem, as well as on the morphological level, enabled the system to better exploit the small number of examples in the corpus. We believe that working with a manually aligned corpus would improve the results dramatically, since we found several cases of perfect fragment matching but bad translation extraction, due to the poor quality of the translation examples. More work is needed to formulate rules to deal with various problematic situations that are not yet handled. This all appears quite feasible.

Finally, we believe that the example-based method is not sufficient to handle the complete translation process. It seems that, for Arabic, at least, it should be combined with some kind of rule-based or statistical-based engine, as part of a multi-engine system, so as to better handle more complicated situations. This is an area of current research. As pointed out above, the key strength of the example-based method is its ability to produce “natural” translations.

References

- Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the 43rd Annual ACL Meeting, Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI, 65-72.
- Brill, Eric. 1992. A Simple Rule-Based Part-Of-Speech Tagger. In *Proceedings of the DARPA, Speech and Natural Language Workshop*. 112-116. Morgan Kaufman. San Mateo, CA.
- Brown, Ralf D. 1999. Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In: *Proceedings of TMI*, 22-32.
- Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. *Linguistic Data Consortium*, Philadelphia, PA.
- Choueka, Yaacov, Ehud S. Conley and Ido Dagan. 2000. A Comprehensive Bilingual Word Alignment System. Application to Disparate Languages: Hebrew and English. In: J. Veronis, *Parallel Text Processing: Alignment and Use of Translation Corpora*, Kluwer.
- Diab, Mona, Kadri Hacioglu and Daniel Jurafsky. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *The National Science Foundation*, Washington, DC.
- Fung, Pascale and Kathleen McKeown. 1994. Aligning Noisy Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping. In: *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-94)*, Columbia, MD, USA, 81-88.
- Habash, Nizar and Owen Rambow. 2005. Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In: *Proceedings of the Conference of American Association for Computational Linguistics*.
- Maruyama, Hiroshi and Hideo Watanabe. 1992. Tree Cover Search Algorithm for Example-Based Translation. In: *Proceedings of TMI*, 173-184.
- Nagao, Makoto. 1984. A Framework of Mechanical Translation between Japanese and English by Analogy Principle. In: A.Elithorn and R.Banerji, eds., *Artificial and Human Intelligence*. North-Holland.
- Nirenburg, Sergei, Stephen Beale and Constantine Domashnev. 1994. A Full-Text Experiment in Example-Based Machine Translation. In: *International Conference on New Methods in Language Processing (NeMLaP)*, Manchester, UK, 78-87.
- Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the ACL 40th Annual Meeting*, Philadelphia, PA, 311-318.
- Phillips, Aaron B., Cavalli-Sforza Violetta and Ralf D. Brown. 2006. Improving Example-Based Machine Translation through Morphological Generalization and Adaptation. In: *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, 369-375.
- Sato, Satoshi and Makoto Nagao. 1990. Toward Memory-Based Translation. In: *COLING 13(3)*:247-252.
- Sumita, Eiichiro and Hitoshi Iida. 1995. Heterogeneous Computing for Example-Based Translation of Spoken Language. In: *Proceedings of TMI*, 273-286.